ISSN 1343-2001

TOKYO UNIVERSITY OF INFORMATION SCIENCES

東京情報大学研究論集

Vol.5 No.1 抜刷

特 集

東京情報大学ハイテクリサーチセンター国際シンポジウム

東京情報大学
2001.8

検索エンジンを利用した日本語Webページ数の統計的推定の研究

水谷正大*、大森貴博**、来住伸子***、小川貴英****

1. 研究の意図

本研究は、ハイテクリサーチ研究において複雑な様相を呈することが期待されている情報処理プロジェクトで構築を目指す情報処理システムの評価研究のための基本研究として、分散システムにおける情報資源の統計的予測を目標として始められた。具体的には分散システムとしてインターネットを、情報資源として日本語World Wide Web（Web） ページを対象とし、この総数を統計的に推定することを研究目標とした。インターネットに代表される分散的に増殖していく情報システムでは、システム全体の状況を把握することは容易なことではない。また、近年のインターネット普及の伸びには目覚しいものがあり、急速に増加しているWebページ数はどのくらい実際に存在するか、また、どの程度のWebページを検索エンジンで実際に検索できるのかについての正確なデータが現状ではほとんど無い。このため、本研究はWeb技術の基礎研究においてWebに関する統計調査方法の研究としても重要である。

2. 研究の概要

1997年にLawrenceらは英語の検索エンジンを使った英語のWebページ数の推定を行ない［1］、英語のWebページ数は1997年には約3億2000万ページあると推定し、当時の最大の検索エンジンでも約34%のページしか検索できないことを指摘した。本研究では日本語の検索エンジンを用いてLawenceと同様な推定方法を使うことにより日本語Webページ数の調査を行ない、日本語Webページ数を英語のWebページ数と比較し、Lawrenceらの方法の仮定するモデルや推定値の統計的信頼性について詳細な検討をすることにした。最近の調査の結果、2000年10月に日本語のWebページは最低約2億5600万個あることが推定できた［23］。また、日本で最大の検索エンジンでも、日本語のWebページ全体の約13%しか検索できないことも分かった。さらに、我々が過去に調査した値も含めると、1999年から2000年にかけて日本語のWebページ数は通信白書で示されている増加より急激に増加していることも分かった。一方、Lawrenceらの推定方法はクエリーの選定等で問題もあり、長期間にわたる調査には適用できない可能性があることが分かった。

本報告では、第2節で従来のWebの調査方法を紹介し、第3節では、Webページ数の推定を中心に、Lawrenceらのモデルにもとづく統計的推定の詳細な説明を行なう。つづいて、第4節では我々が実際に使用した実験方法を述べ、第5節で、その実験結果を紹介し、また日本語Webページの数の推定値の他に、無効URLと無効ページ存在率も紹介する。これらは、推定に使用した検索エンジンが収集しているページがどの程度最新のWebの状態と一致しているかを示す指標となる。第6節では最新の実験結果を報告し、7節では実験結果を考察し、仮定するモデルの妥当性やこの推定方法の

* 東京情報大学 教授
** 東京情報大学大学院経営情報学研究科博士後期過程
*** 津田塾大学数理情報科学科 助教授
**** 津田塾大学数理情報科学科 教授

2001年5月16日受理

妥当性や限界について報告する。

3. 従来の方法と本研究の背景

1989年にWebの構想が提案されて以来、急速にWebサーバー数が増加し始めた［2,3,4］。1993年からの2年間程はWebがまだ普及していなかったため、検索エンジンが実際に集めたWebページ数が世界のWebページ数と考えられた。たとえば、1995年11月にOpenTextは11,366,121個のURLについて数種の統計量を報告している［5］。しかし、Webページ数の増加とともに、単独の検索エンジンではすべてのWebページを直接収集できなくなってきた。

そこで、Bharatらは、複数の検索エンジンを利用し、異なる検索エンジンの検索結果として得られたURL群の中で重複するURLを利用して、Webページの調査を行なうことを提案した［6］。しかし、Bharatらの方法は、重複するURL集合の計算方法等に問題があり、その問題点を改善した方法を提案したのが、Lawrenceらである［1］。Lawrenceらは、米国NECの研究所所員が検索エンジンを利用した履歴から302個のクエリーを選び,それらAltaVista, Excite, HotBot, Infoseek, Lycos, NorthernLightの6検索サービスに与えた。検索結果として返ってきたそれぞれのURL群から、各検索サービス間のURLの重複を計算し、検索可能な公開されたWebページ数が1997年12月には最低3億2000万ページであったと推定した。さらに、1999年2月にも同様の調査を行ない最低3億3500万ページと推定している［18］。また、1999年2月の調査では、IPアドレスのランダム抽出によるWebページ数の推定も行なっており、英語のページに限定しない日本語を含むすべての言語のWebページ数は8億ページだと推定している。

日本では、平成12年版通信白書に1998年2月に1,020万ページ、1998年8月に1,790万ページ、1999年2月に2,950万ページ、1999年8月に3,850万ページという数値が記載されている［9］。これらの数値は、郵政省郵政研究所による調査［7,8］に基づいており、この調査は、実際に収集したWebページ数から線形予測した値を採用している。一方、検索エンジンgooは1998年6月26日に1,700万ページ、1999年11月11日に3,500万ページ［13,14］、検索エンジンLycos日本語版は1999年5月17日に3,000万ページ［15］を収集したことを公表している。これらの数をそのまま解釈すると、日本のWebページの７割以上が単独の検索エンジンで検索できていることになるが、これは、多くのユーザの実感に合わない。したがって、通信白書で示された数値は日本語Webページ数の実数調査が1998年頃には難しくなったことを示している。そこで、本研究ではLawrenceらの推定方法を日本語のWebページに用いることができるかどうか、また、用いることができた場合にどのような推定値を示すかを、実際に調査することにした。

以下の節で見るように、本研究の推定方法が実際に適用できるかどうかは次の条件に大きく依存する。

- 対象とするWebページ集合から、十分に多くのWebページを収集した検索エンジンが複数存在する。
- 適切なクエリーによって十分に大きい標本集団を得ることができる。

4. Web ページ数の推定方法

4.1 対象とするWebページ

Webページの統計的推定に関する本研究では、全文検索システムのクエリーに使用できるような文字列を最低1個は含むテキストデータを標本Webページとし、次のようなWebページは除外する：

- FirewallやWebServerなどのアクセス制限の対象になっている。
- “ robot.txt ” を使って、webロボットが収集しない設定になっている。

これらのWebページ集合を、Lawrenceらはpublicly indexable webと呼んでいる［1］。以降、このURL集合をUとし、$N \equiv |U|$をWebページ数と呼ぶ。

4.2 用語の定義

いま、次の集合を考える：

S : 検索エンジンの集合

Q : クエリーの集合

検索エンジンの集合Sとは、goo、lycosなど存在するすべての検索エンジンで有限集合である。クエリーの集合Qとは、これらの検索エンジンに与えることのできる検索文字列の集まりである。語を任意個数並べてよいと考えると、事実上Qは無限集合である。

これらの集合の要素$s \in S, q \in Q$を使って、次の2種類のURLの集合を定義する:

$$U_s{}^q \equiv \{u \mid \text{クエリー } q \text{ を検索エンジン } s \text{ に与えると得られる検索結果中の URL } u\}$$

$$U_s \equiv \bigcup_{q \in Q} U_s{}^q.$$

つまり、Usは検索エンジンsによって検索可能なURL集合である。

4.3 確率の定義

全URL集合Uの任意の部分集合に対して確率$P(X)$を次のように定義する:

$$P(X) \equiv \frac{|X|}{|U|}.$$

このとき、事象Usである確率

$$P(U_s) = \frac{|U_s|}{|U|}$$

と検索エンジンsが持つURL集合の大きさ$|Us|$が分かれば、Webページ数$N=|U|$は

$$N = \frac{|U_s|}{P(U_s)}$$

で求めることができる。

4.4　確率の推定方法

2つの事象A, Bが独立であれば、

$$P(A) = P(A|B) = \frac{|A \cap B|}{|B|}$$

が成り立つ。2つの検索エンジンaとbのWebクローラーが互いに独立してWebページを収集していると仮定すると、U_aとU_bは独立した事象と見なせるので、次が成立する：

$$\begin{aligned} P(U_a) &= P(U_a|U_b) \\ &= \frac{|U_a \cap U_b|}{|U_b|}. \end{aligned}$$

このとき、P（U_a）は｜U_a∩U_b｜/｜U_b｜から求めることができる。しかし、検索エンジンは収集ページ数｜U_a｜や｜U_b｜を公開しても、その内容U_aやU_bを通常公開しない。そのため、実際にU_a∩U_bを調べることは難しい。そこで、次のように工夫してP（U_a）の近似値を求めることにする。

クエリーの有限部分集合Q'⊂Qを選び出して

$$U'_a \equiv \bigcup_{q \in Q'} U_a^q$$

を定義する。U'_aは、有限個のクエリーを用意すれば、実際に検索エンジンを使って観察することができる。また、Q'がQからランダムに選ばれるならば、近似的に次の関係がなりたつと期待できる。

$$\frac{|U_a \cap U_b|}{|U_b|} \approx \frac{|U'_a \cap U'_b|}{|U'_b|}.$$

このとき、P（U_a）は次のように推定できる：

$$P(U_a) \approx \frac{|U'_a \cap U'_b|}{|U'_b|}.$$

これを式(1)に使ってNを得る。また、P（U_b）からも、同様にNを得ることができる。そこで、2つのNの平均を、検索エンジンaとbから得られるNの推定値とする。

4.5　区間設定

Nページの中から非復元的にランダムにn=｜U'_b｜ページ取り出したとき、nのうちのXページがU_aに属するXの確率分布は超幾何分布となる［16,p.109-111］。n≪Nの場合、1ページを取り出す結果は次の1ページを取り出す結果にほとんど影響しない。したがって、p =P（U_a）とおくと、Xの確率分布は2項分布$B(n, p)$となる。このとき、期待値E（X）と分散σ（X）は、次のようになる。

$$E(X) \approx np$$

$$\sigma^2(X) \approx np(1-p)$$

nが大きいと、中心極限定理により、f（x）は、正規分布に近付づく［16,p.170］。 $\bar{p} = \frac{|U'_a \cap U'_b|}{|U'_b|}$ をpの観測値とすると、pの95%信頼区間は近似的に

$$\left[\bar{p} - 1.96\sqrt{\bar{p}(1-\bar{p})/n},\quad \bar{p} + 1.96\sqrt{\bar{p}(1-\bar{p})/n}\right]$$

で求めることができる。

5. 実験方法

本研究では、次の3検索エンジンを今回の調査に使用することにし、以下にのべる手順で実験を行なった：

- goo［10］
- Lycos日本語版［11］
- Infoseek日本語版［12］

3検索エンジンの他に、AltaVista,Ring,Excite日本語版なども調査に利用することを検討したが、検索結果数が安定しない、無効URL率が高い、収集Webページ数を公表していないなどの理由で推定には利用しなかった。

手順1：URL集合の取得とクエリーの選定　まず、1994,95年の2年間の毎日新聞記事を形態素解析し、英数字を含まず、ひらがな、カタカナ、漢字のどれか一種からなる名詞、約143,461語を選んだ。英数字を含む語を除いた理由は日本語を含むページを検索対象とするためであり、かな漢字混じり語を除いた理由は検索エンジンによる語の分割の可能性を低くするためである。

このようにして選んだ語を、上記の3検索エンジンで検索し、その検索結果としてURL集合を得る。さらに、このURL集合を調査し、次の3条件にあてはまる291語を調べ、これを有限クエリ集合（Q'）として採用した：

a　各検索エンジンでの検索結果のURL数が50個以上である。
b　3検索エンジンの検索結果の和集合の大きさが600個以下である。

条件aの下限は、ある検索エンジンでインデックスの対象に全くならないクエリーを採用しないために設定している。50という値はLawrenceらの調査と一致させるために採用した。条件bは、単独のクエリーの検索結果が大きな影響を与えない役割を果たしている。また、条件bにより、各検索エンジンでの検索結果件数も600個以下になるが、Infoseekは表示URL数の上限から500個以下にしている。

手順2：URLの正規化　次に、上記のURL集合の各要素であるURL名に以下の正規化を行ない、重複するURL名を除外した：

1.ホスト名の小文字化
2.ホスト名の後の:80の除去
3.16進数表現文字の変換（例:%7Eを~ にする）
4.index.htmlの除去して/で終わるURLとして統一する。

手順3：ページ集合の取得　上記で得られたURL全てについて、httpGET要求を出すことにより、そのURLに対応するWebページが実際に存在するかを調べた。調べた期間は2000年10月1日から3日の間で、ページの存在が確認できなかったURLと確認時にTimeoutしたURL（今回の実験では80秒）を無効URLとしてURL名の集合から除いた。

手順4：クエリーの存在確認　GET要求で取得できたWebページの内容に、クエリーに該当する文字列が含まれるかどうかを調べた。perlの文字列パターンマッチ機能を使い、空白文字、改行文字と中黒「・」を含むクエリーがあればクエリーを含むページとし、そうでない場合はクエリーを含まないページとした。

手順5：重複ページ集合の生成　最後に、3検索エンジンから互いに異なる2検索エンジンを選び、それら3組について重複ページ集合$U'a \cap U'b$を求めた。また、3検索エンジンで検索できたページ集合の和集合も求めた。

6. 実験結果

表1：日本語Webページ数の推定

a	b	$\lvert U'a \rvert$	$\lvert U'b \rvert$	$\lvert U'a \cap U'b \rvert$	$\overline{p}=P(Ua)$	$\overline{p}$の95% 信頼区間	N
Goo	Lycos	17,125	44,817	4,047	0.090	0:088 〜 0:093	256; 000; 000
Lycos	Goo	44,817	17,125	4,047	0.236	0:230 〜 0:243	
Goo	Infoseek	17,125	65,893	5,721	0.087	0:085 〜 0:089	230; 000; 000
Infoseek	Goo	65,893	17,125	5,721	0.334	0:327 〜 0:341	
Lycos	Infoseek	44,817	65,893	21,890	0.332	0:329 〜 0:336	59; 300; 000
Infoseek	Lycos	65,893	44,817	21,890	0.488	0:484 〜 0:493	

6.1　実験結果1：日本語Webページ総数

実験の手順5で生成した重複集合を利用して、最新（2000年10月13日）の日本語Webページ総数Nについて、表1に示す推定値を得た。この表では、3検索エンジンが公表した収集ページ数をもとにNの推定を行なっている。goo［14］は3500万ページ、lycos［15］は3000万ページ、infoseekは1800万ページと収集ページ数を公表している。

表1のNの欄で示す推定値が一致しない原因はいくつかある。まず、重複集合の占める割合$\overline{p}$にばらつきがある。これは、クローラーの収集するページ集合の独立性にばらつきがあることを示す。

つぎに、検索エンジンが公表したURL数の範囲の違いが考えられる。たとえば、クローラーが収集したページ全てを数えるか、クエリーに利用できるような文字列を最低１個含むページだけを数えるかで収集したページ集合の大きさが異なる［17］。

最新の調査では、Lawrenceらの方法にならって、公表したページ数の大きい組み合わせであるgooとlycosから推定した、256×10^6を日本語Webページ数として採用することにした。第6節で、この組み合わせを元に推定する理由についてさらに考察する。

6.2 実験結果2: 無効URLと無効ページ

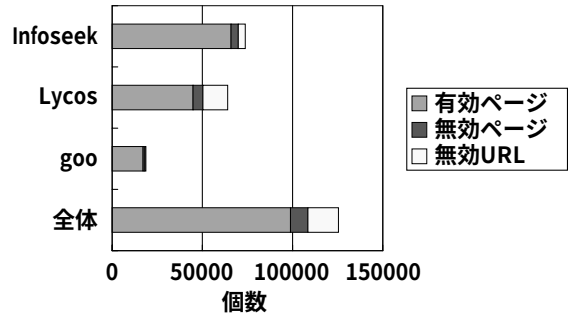

図1：各検索サービスにおける無効ページと無効URL数

重複URL集合を計算する前の段階、手順3で除いたURLは次のようなものであり、これらを無効URLと呼ぶ。

（a）connectできない、80秒でTime outしたなどの理由で、Webサーバーの存在が確認できなかった。

（b）ページの存在が確認できなかった。

また、手順4で除いたページは、ページは存在したがクエリーを含まなかったページで、これらを無効ページと呼ぶ。無効ページの原因は、大まかには次の二つである。

（c）新聞記事、電子掲示板、日記など、頻繁に書き換えられるページであるため、検索エンジンがインデックスを生成した時と内容が異なる。

（d）クエリーを分割して複合語検索をしたり、関連語も検索対象にしたりしたため、単純なパターンマッチでクエリーの存在を確認できなかった。

無効URLや原因（c）の無効ページが多いのは、検索エンジンのWebクローラーの収集頻度が低いためと考えられる。そこで、無効URLが元のURL群に占める割合を無効URL存在率、無効ページが元のURL群に占める割合を無効ページ存在率とし、表2と図1に示す。なお、今回の調査では、無効ページの原因（c）と（d）を区別していない。

この表から、無効URL存在率と無効ページ存在率の両方ともgooが最も低く、もっとも最近に収集を行なった可能性が高いことが分かる。また、他の2検索エンジンでも、無効URL率と無効ページ存在率を合わせて2割程度で、Webの更新状況を比較的よく反映している［19］。

この表は、使用したクエリー集合に対する検索結果件数は、Infoseekが最も大きいことも示している。これは3検索エンジンの中でInfoseekの収集ページ数が最も小さいこととは矛盾しておらず、Infoseekは検索結果件数が500個以下の範囲におさまるクエリーの存在確率が、他の2検索エンジンより高いことが原因だと考える。つまり、第4節で示した条件bを適用せず、600個以上の検索結果件数になるクエリーも含めた調査ができれば、実際に収集したページ数やインデックスの大きさに近い大小関係になっただろうと考える。

表2：各検索サービスにおける無効URL存在率

	検索ページ数	有効ページ数	無効ページ数	無効ページ存在率(%)	無効URL数	無効URL存在率(%)
Infoseek	73771	65893	3999	5.4	3879	5.2
Lycos	64932	44817	5519	8.6	13696	21.3
goo	18674	17125	926	4.9	623	3.3
全体	125423	98804	9711	7.7	16980	13.5

7. 考察

7.1 推定値とモデルの妥当性

本研究では検索エンジンの検索結果の重複を利用した推定方法を日本語Webページに適用し、日本語Webページ数を推定した。Lawrenceらの仮定したモデルが正しいとすると、2000年10月の時点で、日本語のWebページ数は2億5600万で、95%信頼区間は±600万ページ程度である。95%信頼区間の幅から、Lawrenceらが行なった調査程度には、十分な大きさの標本集合を調査したことが分かる。

しかしながら、実際のWebページ集合にLawrenceらの仮定したモデルがあてはまらなくなる要因として次のものが考えられる。これらの要因の影響について順に考察する。

1. 検索エンジンのクローラーの動作の独立性
2. 検索エンジンのインデックスの作成方法
3. 検索エンジンが公表したURL数の範囲

7.1.1 クローラーの動作の独立性

検索エンジンがWebページを収集するとき、つぎのようなページを優先して集めることがある。

- ユーザが検索対象にしてほしいと検索エンジンに登録したページ。
- 人気のあるWebサイト（例：Yahooディレクトリ）からリンクされているページ。
- 被リンク数の多いページ。特に、そのページの属するWebサイト以外からの被リンク数の多いページ。
- URL名のパス名部分が短い（Webサーバーのルートドキュメントに近い）ページ。
- 更新頻度が高いページ。

このようなページを優先して集めた場合、異なる検索エンジンが収集したページ集合の間の重複集合は、ランダムに集めた場合の重複集合より大きくなる。しかし、上記の条件にあてはまるページ数は限られているで、検索エンジンが集めたページが多くなるほど、優先して集めたページが重複集合の中に占める割合が小さくなると考えられる。つまり、検索エンジンが集めたページ数が大きくなるほど、片寄りのある場合から、ランダムに集めた場合に近づくと考えられる。実際、Lawrenceらの調査でも我々の調査でも、検索結果集合の中に占める重複集合の大きさの割合は、公表した収集ページ数が大きい2検索エンジンの間のものが最も低い。

そこで、我々の調査でも公表した収集ページ数の大きいgooとLycosの値を利用することにした。仮にこの二組の検索エンジンがページを互いに独立して収集していないとしても、Webページ数の推定値を小さくする方向に影響するので、クローラーの独立性は、「Webページ数の下限」の推定には影響しないと考える。

7.1.2 インデックスの作成方式

収集したページ集合が同じでクエリーが同一であっても、インデックスの作成方式が同じでないと検索結果が異なることがある。

第4節の条件aにより各検索エンジンでの検索結果のURL数がゼロでないことが保証されるので、

検索エンジンのどれかが完全に無視する文字列を除外している。

条件bでは和集合の大きさに制限を加えて、各エンジンのクエリーの取り扱いが大きく異なる場合をある程度除外している。それでも、使用したクエリーには、表記のゆれの影響をうける語がいくつかある。これは、4節の手順1で複合語検索を起こさないためにつけた制約によって、クエリーにカタカナ語が多くなったためと考えられる。たとえば「キャフェ」で検索した場合、「カフェ」を含むページのURLが検索結果に含まれる可能性がある。そこで、表記のゆれや関連語への対応をしている検索エンジンの検索結果には、元のクエリーを含まないページが多数含まれる可能性がある。検索結果のURLに対応するページの内容にクエリーに使用した文字列自体を含むことを調べ、含まれない場合は無効ページとして、Webページ数の推定には利用していない。また、無効ページ数が検索結果に含まれる割合は、表2と図1に示すように、最も高い検索エンジンでも8.6%であった。そこで、今回の調査結果は関連語や表記のゆれの影響を受けていたとしても、約10（≈8.6=91.4）%程度であろうと考える。

一方、表記のゆれや関連語検索とは逆に、あるクエリーqがページpに含まれているにも関わらず、クエリーqの検索結果にpが含まれていないことがある。たとえば、我々が1999年10月の調査で使用した597語を並べたページをWeb上で公開しておいたところ、ある検索エンジンのクローラーで収集され、その検索エンジンで実際に検索できるようになった。しかし、597語全部では検索はできず、約240語で検索できただけであった。このように検索できないクエリーqが存在する理由は、次のようなものが考えられる：

- クエリーqより重要な検索語がページpに非常に多く含まれていたために、qがインデックスの対象から除外された。
- 検索エンジンがページの先頭から一定の範囲内にある語だけをインデックスの対象にしており、クエリーqがページの後ろの方にあるために、インデックスの対象から除外された。

このようなクエリーqが、今回の調査で使用したクエリーの中に多く存在するとある検索エンジンでは収集されているにも関わらず、検索結果に反映されないページが多くなる。そのため、調査した重複集合の大きさが実際の重複集合の大きさより小さくなり、Webページ数の推定数が実際より大きな数になる。そこで、このようなページ、重要度の低いクエリーqを使用したために検索できなかったページは少ないことを次の方法で確認することにした。

gooとLycosの検索結果から、片方のみから検索できたページで、サイズの大きいページを10個選ぶ。次に、そのページを眺め、そのページに含まれる語でクエリーに使用するのが妥当だと思われる語を選び、それらを使って検索し、検索結果件数が600件以下の5語を選んだ。5語の検索結果に、10個のページは含まれていなかった。つまり、検索できなかった方の検索エンジンから新しく検索できるようになったページはなかった。

本来は、上記の方法ではなく、今回の調査で標本集合として採用した約9万8000ページについて、各検索エンジンで収集されているかいないかを調査すべきだが、各検索エンジンはそのような手段を提供していないので、上記の方法で代用した。

7.1.3　公表しているページ数の範囲

このWebページ数の推定では、検索エンジンが公表している収集ページ数を利用しているので、

収集ページとは何を意味しているによって日本語Webページ数の範囲が異なる。通常、クローラーが収集したページには、インデックスの対象にならない次のようなページが含まれていることが多い。

- ファイルとしては存在するが、タグのみを含み、テキストを含まないページ
- テキストを含むが“Under Construction”のような非常によく使われるテキストであるためインデックスの対象とならない語のみを含むページ
- 日本語と英語以外のテキストのみを含むページ
- すでに収集したページと内容が同一とみなせる（ミラー）ページ

各検索エンジンが、上記のページを含んだ数をページ数として公表しているのであれば、我々の推定した数は、上記のページを含んだ数になる。実際には、各検索エンジンは公表ページ数を「登録URL件数」、「日本語サイト3500万URL」などと記しているので、上記のページはある程度除き、データベースに登録できたページ数を公表していると考える。一方、通信白書の採用した方式では、クローラーで集められるページをすべて数えており、今回の調査対象であるページ集合より広い範囲のページを含むページ集合を調査した推定値であると考える。

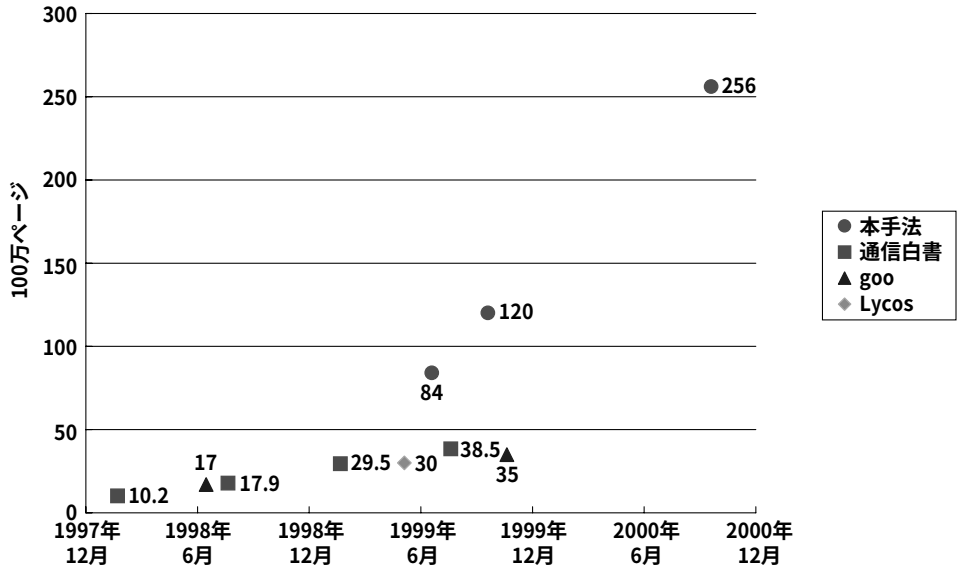

図2：日本語Webページ数の推定数の変化

7.1.4　3要因の影響

上記の3要因が強く影響をあたえると、本研究で調査した推定値は、Webページ数の下限値として小さすぎる可能性と大きすぎる可能性の両方がある。

次の要因が強く影響を与えていると、小さすぎる推定値になる。

- クローラーの動作に独立性がない。

●検索エンジンが実際より少ない収集ページ数を公表している。

この研究で推定しようとしているのは日本語Webページ数の「下限」なので、小さすぎる推定値になるのは止むを得ないと考える。

一方、次の要因が強く影響を与えていると、大きすぎる推定値になる。

●インデックスの対象にならないクエリーを多く使用した。

●検索エンジンが実際より大きい収集ページ数を公表している。

最初の要因によって大きすぎる推定値である場合は1,2割程度の過大評価であり、その分を考慮しても、通信白書の推定値よりかなり大きいといえる。図2は、我々の推定値と通信白書の推定値の関係を示す。この図から、日本語のWebページ数は通信白書の示す値以上に急速に増大していることが分かる。

7.2　経時変化の測定

今回、我々がLawrenceらの方法を日本語Webに適用できた理由は、彼らが使用した条件と同じ条件を満たす日本語のクエリーを十分な数見つけることができ、その結果十分な量の標本集合を集めることができたためである。

推定時期	日本語Web ページ 推定数 N	使用した クエリー数	クエリーの出典
1999 年 7 月	84×10^6	1085	検索エンジンの利用履歴
1999 年10月	120×10^6	597	現代用語の基礎知識 99の見出し語
2000 年10月	256×10^6	291	毎日新聞記事（1994、95）

まず、第1の問題点として、調査に利用できる日本語検索エンジンの少なさがあげられる。この調査に利用する日本語検索エンジンは次のような条件を満たす必要がある。

●収集ページ数が大きい。

●収集ページ数を公表している。

●更新頻度が高い。

●サービスを安定して提供している。

この条件にあてはまる検索エンジンは今回の調査ではLycos, goo, Infoseekの3サービスしかなく、そのうち、検索結果件数に制限がないのは2サービスであった。今後、上記の条件に適合する日本語検索エンジンが多く出現することが期待される。

第2の問題点としてクエリーの選定の難しさがあげられる。我々はこの手法を日本語のWebに過去数回適用してきた［20,21］。その度に条件をみたすクエリー集合を変更した。表3は、手法の適用時期と、使用したクエリー集合、推定したNの大きさを示す。表3は、日本語のWebページ数が増加と共に、クエリー集合が小さくなり、時期ごとにクエリー集合の出典が異なることを示している。これは、ある調査であるクエリーが検索結果件数の和集合の大きさ600以下の条件を満たしても、次の調査では600を超えることが多く、クエリーに要求される条件を満たさなくなる。つまり、現在の方法は調査のたびにクエリーを毎回選定しなおす必要があり、将来も適切なクエリーを見つけ

出せるだせるかどうか分からない。この問題点を解決するには、600以下の条件をゆるめて600より多くの検索結果件数になるクエリーを使用することが考えられる。一方、第1の問題点で指摘したように、この手法に利用可能な、600より多くの検索結果件数を得ることのできる日本語検索エンジンは現状では2サービスしかない。調査に利用できるクエリー数の拡大には日本語検索エンジンの充実が望まれる。

第3の問題点として、検索エンジンの設計方針の変化の影響が考えられる。現在、英語圏の検索エンジンは大量の検索結果を表示するのではなく、少数でも有用な検索結果を表示する方向への変化が始まりつつある。たとえば“Internet”のような非常に多くのページに含まれるクエリーについては同一サイトからは一定数のページしか表示しない、かなり広い範囲のページをミラーとみなして除外するという工夫が行なわれつつある。今回の調査に使用した日本語の検索エンジンではそのような傾向を観察できなかったが、この傾向が強くなると、検索エンジンにクエリーを与えて標本集合を得る考え方自体が推定に向かなくなる。

8. まとめ

Lawrenceらの方法［1］を日本語Webに適用した結果、日本語Webには2000年10月には、少なくとも約2億5600万ページ存在することが推定できた［23］。また、日本で最大の検索エンジンでも日本語Webページの約13%しか検索できないことも推定でき、1997年12月の英語Webページ検索の状況より悪化している。一方、Lawrenceらの検索エンジンを使った推定方法は、長期間にわたる経時変化の測定や世界全体のWebページ数の推定にあまり適していないことも分かった。

なお、検索エンジンを利用した推定方法の限界を解決するために、Lawrenceら自身も検索エンジンを利用しないWebページ数の推定方法を後日発表し、1999年2月の世界のWebページ数は8億個であるとした［18］。この方法はIPアドレスをランダムに抽出して行なうものだが、問題点がいくつかある。たとえば、JPドメインに同じ方法を適用してみると、約1900万ページになる［22］。また、複数ドメイン名の同一IPアドレスへの割り当て、IPv6の導入などの影響により、国別や言語別のIPアドレス分布は今後大きく変化すると予想されるので、長期的な観測にはLawrenceらの新しい方式もあまり適していない。Webに関する統計的推定には、さらに新しい手法の開発が今後必要と考えられる。

参考文献

［1］Lawrence, S. and Giles, C.L.: Searching the World Wide Web, SCIENCE 280 , pp.98 100（1998）. http://www.sciencemag.org/cgi/content/abstract/280/5360/98 , http://www.neci.nj.nec.com/homepages/lawrence/websize.html
［2］Robert Cailliau: A Little History of the World Wide Web. http://www.w3.org/History.html
［3］Mathew Gray: Measuring the Growth of the Web – June 1993 to June 1995 http://www.mit.edu/people/mkgray/growth/
［4］Berners-Lee, T.: Weaving the Web: The Original Design and Ultimate Destiny of theWorld Wide Web by Its Inventor, HarperCollins（1999）.
［5］Bray, T.: Measuring the Web Fifth International World Wide Web Conference（1996）. http://www5conf.inria.fr/fich html/slides/papers /PS3/P9/T01.htm
［6］Bharat, K. and Broder, A.: A technique for measuring the relative size and overlap of public Web search engines, Seventh Iternational World Wide Web Conference,（1998）.

http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm
［7］外薗博文: 日本のインターネット（WWW）の現状, 郵政研究所月報9 , pp.79　86（1998）.
［8］宮沢浩: 日本のインターネット（WWW）の現状その2, 郵政研究所月報12 , pp.99　102（1998）.
［9］郵政省: 平成12年版通信白書
http://www.mpt.go.jp/policyreports/japanese/papers/h12/ html/C1Z10000.html
［10］http://www.goo.ne.jp
［11］http://www.lycos.co.jp
［12］http://www.infoseek.co.jp
［13］エヌ・ティ・ティ・アド：350 万ページビュー／１ 日突破　http://www.goo.ne.jp/help/n_980626.html
［14］エヌ・ティ・ティエムイー情報流通: ポータルサイト“ goo ”の検索機能を大幅に強化
http://www.goo.ne.jp/help/info/n_release/n_001005.html
［15］Lycos Japan: Lycos Japan がサイト全般に渡る大規模リニューアルを実施
http://www.lycos.co.jp/help/info/release.html?press=06
［16］東京大学教養学部統計学教室編: 統計学入門| 基礎統計学I, 東京大学出版会(1991).
［17］How to count URLs
http://www.excite.com/ice/counting.html
［18］Lawrence, S. and Giles, C.L.: Accessibility of information on the web, NATURE 400 , pp.107{
109 (1999). http://www.wwwmetrics.com
［19］Brewington, B.E. and Cybenko,G.: How dynamic is the Web? Proceedings of 9th International World Wide Web Conference, pp. 257{275 (2000).
［20］大森貴博, 笹塚清二, 近藤晶子, 水谷正大, 来住伸子, 小川貴英: 統計的手法による日本語Web の調査、情報処理学会第59 回全国大会（平成11 年後期）講演論文集、3P-01, (1999).
［21］来住伸子, 大森貴博, 笹塚清二, 近藤晶子, 水谷正大, 小川貴英：統計的推定による日本語Web の調査, インターネットコンファレンス'99 論文集、日本ソフトウェア科学会研究会シリーズNo.14, ppp.21-28 (1999).
［22］Kishi, N., Ohmori, T., Sasazuka, S., Kondo, A., Mizutani, M., and Ogawa, T.: Estimating Web Properties by Using Search Engines and Random Crawlers, INET 2000 Proceedings, Ineternet Society, (2000).
http://www.isoc.org/inet2000/cdproceedings/2a/2a 3.htm
［23］来住伸子，大森貴博、水谷正大、小川貴英：検索エンジンを利用した日本語Web ページの統計的推定、IPSJ Symposium Series: データベースとWeb 情報システムに関するProceedings of DBWeb2000、Vol.2000 No.14, pp149-156 (2000)

ISSN 1343-2001

Journal of Tokyo University of Information Sciences

Reprinted from Vol.5 No.1

Symposium

TOKYO UNIVERSITY OF INFORMATION SCIENCES

2001.8
Published by Tokyo University of Information Sciences